A Probability Distribution Associated With Events With Multiple Occurrences

Panaretos, John and Xekalaki, Evdokia

1989

A PROBABILITY DISTRIBUTION ASSOCIATED WITH EVENTS WITH MULTIPLE OCCURRENCES

John PANARETOS

University of Patras, Greece

Evdokia XEKALAKI

Athens University of Economics, Greece

Received August 1988
Revised October 1988

1. Introduction

In a recent paper Xekalaki and Panaretos (1989) introduce what they call the cluster negative binomial and cluster negative multinomial distributions in the context of cluster sampling. Clusters are represented by balls each marked with a number in the range $\{0, 1, 2, \ldots, k\}$ where k is a positive integer. Balls are drawn one at a time and with replacement till a given number of zeroes is observed. The distribution of the total sum of the sampled numbers is defined to be the cluster negative binomial distribution. The case of straight sampling was considered by Panaretos and Xekalaki (1986a). Given a sample of n balls (clusters) numbered as previously and drawn with replacement the sum of the sampled numbers was shown to have a probability distribution termed the cluster binomial distribution. As mentioned by Xekalaki and Panaretos (1989), such urn schemes may be useful in modelling random mechanisms in the area of biology where the frequency distribution of cells forming clusters is of interest. Epidemiology is another area of possible application. There, the hypothesis of contagion is very common in determining the probability distribution of infected cases during an epidemic. In sampling-from-an-urn schemes this hypothesis can be accounted for by additional replacements which alter the structure of the urn. Panaretos and Xekalaki's (1986a) urn scheme giving rise to the stuttering generalized Waring distribution is pertinent to such a situation. There, balls marked $0, 1, 2, \ldots$, or k are sampled with additional replacements till a given number of zeroes is observed. Of course among the various investigators who have used urn models for describing contagious biological data Polya (1931) seems to be a pioneer in the field.

This paper considers again sampling from an urn containing balls marked with an integer in the range $\{0, 1, 2, \ldots, k\}$, $k > 0$, and studies the probability distribution of the sum of numbers shown on a sample of n balls drawn without replacement or with additional replacement(s). As a result, hypergeometric and

negative hypergeometric types of distributions are obtained. Their structural and distributional properties are studied and their association to other distributions arising under various schemes for sampling clusters are examined.

2. The cluster hypergeometric distribution

Consider an urn filled with α balls bearing number 0 and β_i balls bearing number i, $i = 1, 2, \ldots, k$. A random sample of n balls is drawn without replacement. If X represents the sum of the numbers shown on the sampled balls then the following theorem can be shown.

Theorem 2.1. *The probability of the event* $\{X = x\}$ *is given by*

$$P(X = x) = \sum_{\Sigma_{i=1}^{k} i x_i = x} \binom{n}{x_1, x_2, \ldots, x_k, n - \Sigma_{i=1}^{k} x_i} \frac{\beta_1^{(x_1)} \cdots \beta_k^{(x_k)} \alpha^{(n - \Sigma_{i=1}^{k} x_i)}}{\left(\alpha + \Sigma_{i=1}^{k} \beta_i\right)^{(n)}}, \qquad (2.1)$$

$x = 0, 1, 2, \ldots, nk$, *where the symbol* $\gamma^{(r)}$ *stands for* $\Gamma(\gamma + 1)/\Gamma(\gamma - r + 1)$, $r = 0, 1, \ldots, \gamma$, $\gamma > 0$.

Proof. Let X_i denote the number of balls bearing number i in the sample of n balls. Then obviously $\{X = x\}$ is the union of all the outcomes $\{X_1 = x_1, \ldots, X_k = x_k\}$ for which $\Sigma_{i=1}^{k} i x_i = x$. i.e.

$$\{X = x\} = \bigcup_{\Sigma_{i=1}^{k} i x_i = x} \{X_1 = x_1, \ldots, X_k = x_k\}.$$

Then (2.1) follows as a consequence of the fact that the vector $(X_1, X_2, \ldots, X_k)$ follows the multivariate hypergeometric distribution with parameters $n, \alpha, \beta_1, \ldots, \beta_k$. $\square$

Theorem 2.2. *The function defined by* (2.1) *is a proper probability function.*

Proof. It is sufficient to show that $\Sigma_{x=0}^{nk} P(X = x) = 1$. Indeed

$$\sum_{x=0}^{nk} P(X = x) = \sum_{x=0}^{nk} \sum_{\Sigma_{i=1}^{k} i x_i = x} \binom{n}{x_1, x_2, \ldots, x_k, n - \Sigma_{i=1}^{k} x_i} \frac{\beta_1^{(x_1)} \cdots \beta_k^{(x_k)} \alpha^{(n - \Sigma_{i=1}^{k} x_i)}}{\left(\alpha + \Sigma_{i=1}^{k} \beta_i\right)^{(n)}}.$$

Letting $x_i \rightarrow x_i$ and $x \rightarrow x + \Sigma_{i=1}^{k}(i - 1)x_i$ we obtain

$$\sum_{x=0}^{n} P(X = x) = \sum_{x=0}^{n} \binom{n}{x} \frac{\alpha^{(n-x)}}{\left(\alpha + \Sigma_{i=1}^{k} \beta_i\right)^{(n)}} \sum_{\Sigma_{i=1}^{k} x_i = x} \binom{x}{x_1, \ldots, x_k} \beta_1^{(x_1)} \cdots \beta_k^{(x_k)} = 1.$$

Hence the theorem has been established. $\square$

Note. One can easily observe that for $k = 1$ the probability distribution defined by (2.1) reduces to the ordinary hypergeometric distribution. Moreover, by the definition of the descending factorial the definition of (2.1) can be extended to include the case of non integer values for the parameters α and β_i, $i = 1, 2, \ldots, k$.

Definition 2.1. Let n, k be fixed positive integers and let α, β_i, $i = 1, 2, \ldots, k$ be positive real numbers. A random variable X will be said to have the cluster hypergeometric distribution with parameters k, n, α, $\beta_1, \ldots, \beta_k$ ($X \sim H_k(n; \alpha, \beta_1, \ldots, \beta_k)$) if its probability function is given by (2.1).

Steyn's (1956) univariate factorial multinomial distribution of the number of occurences of an event E when sampling from a population whose items cause multiple occurrences of E can also be obtained from (2.1) if the proportion of balls marked i is allowed to be an arbitrary real number in (0, 1) instead of a rational number in (0, 1).

Theorem 2.3. *Let X be a non-negative integer-valued random variable having the cluster hypergeometric distribution with probability function as given by (2.1). Then*

(i) $$G_X(s) = \frac{\alpha^{(n)}}{\left(\alpha + \Sigma_{i=1}^k b_i\right)^{(n)}} F_D\left(-n; -\beta_1, -\beta_2, \ldots, -\beta_k; \alpha - n + 1; t, t^2, \ldots, t^k\right) \tag{2.2}$$

where $G_X(s)$ denotes the probability generating function of X, and F_D is Lauricella's generalized hypergeometric function derived by

$$F_D(a; b_1, \ldots, b_k; c; z_1, \ldots, z_k) = \sum_{r=0}^{\infty} \sum_{\Sigma_{i=1}^k r_i = r} \frac{a_{(\Sigma_{i=1}^k r_i)}(b_1)_{(r_1)} \cdots (b_k)_{(r_k)}}{c_{(\Sigma_{i=1}^k r_i)}} \frac{z_1^{r_1}}{r_1!} \cdots \frac{z_k^{r_k}}{r_k!} \tag{2.3}$$

(here $a_{(r)}$ stands for $\Gamma(a + r)/\Gamma(a)$, note that when $a < 0$ the series is terminating, i.e. r runs in the range $0, 1, \ldots, [a]$);

(ii) $$E(X) = \frac{n}{\alpha + \Sigma_{i=1}^k \beta_i} \sum_{i=1}^k i\beta_i; \tag{2.4}$$

(iii) $$V(X) = \frac{n\left(\alpha + \Sigma_{i=1}^k \beta_i - n\right)\left\{\Sigma_{i=1}^k i^2\beta_i\left(\alpha + \Sigma_{j \neq i}^k \beta_j\right) - 2\Sigma_{i<j} ij\beta_i\beta_j\right\}}{\left(\alpha + \Sigma_{i=1}^k \beta_i\right)^2\left(\alpha + \Sigma_{i=1}^k \beta_i - 1\right)}. \tag{2.5}$$

Proof. (i) Applying the transformation $x_i \to x_i$, $x \to x + \Sigma_{i=1}^k(i-1)x_i$ we have

$$G_X(s) = \sum_{x=0}^n \sum_{\Sigma_{i=1}^k x_i = x} \binom{n}{x_1, \ldots, x_k, n-x} \frac{\beta_1^{(x_1)} \cdots \beta_k^{(x_k)}\alpha^{(n-x)}}{\left(\alpha + \Sigma_{i=1}^k \beta_i\right)^{(n)}} s^{\Sigma_{i=1}^k ix_i}$$

$$= \frac{\alpha^{(n)}}{\left(\alpha + \Sigma_{i=1}^k \beta_i\right)^{(n)}} \sum_{x=0}^n \sum_{\Sigma_{i=1}^k x_i = x} \frac{\beta_1^{(x_1)} \cdots \beta_k^{(x_k)}(\alpha - n)^{(-\Sigma_{i=1}^k x_i)}}{(1-n)^{(-\Sigma_{i=1}^k x_i)}} \prod_{i=1}^k \frac{s^{ix_i}}{x_i!}$$

which is equivalent to (2.2).

The proofs of parts (ii) and (iii) follow as a result of the fact that $(X_1, X_2, \ldots, X_k)$ has the multivariate hypergeometric distribution with $E(X_i)$ and $E(X_iX_j)$ given by the appropriate well-known formulas for $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, k$. $\square$

As is well known in the context of urn models the ordinary hypergeometric distribution is the analogue of the binomial distribution when sampling is done without replacement and the former has the latter as a limiting form when the numbers of balls are increased so that their proportion tend to some value $p \in (0, 1)$. It was mentioned in the introduction that a scheme analogous to that considered in this section but whereby sampling is done with replacement gives rise to Panaretos and Xekalaki's (1986b) cluster

binomial distribution. This is the distribution defined by the probability function

$$P(X=x) = \sum_{\Sigma_{i=1}^{k} i x_i = x} \binom{n}{x_1, x_2, \ldots, x_i, n - \Sigma_{i=1}^{k} x_i} q^{n - \Sigma_{i=1}^{k} x_i} \prod_{i=1}^{k} p_i^{x_i}, \tag{2.6}$$

$q = 1 - \Sigma_{i=1}^{k} p_i$, $0 < p_i < 1$, $i = 1, 2, \ldots, k$, $x = 0, 1, \ldots, nk$.

It would therefore be interesting to examine whether the distribution in (2.6) can serve as an approximation of (2.1). The following theorem shows that this is indeed the case.

Theorem 2.4. *Let X, Y be two random variables whose probability functions are given by (2.1) and (2.6) respectively. If $\alpha \to \infty$, $\beta_i \to \infty$, $i = 1, 2, \ldots, k$ so that $\beta_i/(\alpha + \Sigma_{i=1}^{k}\beta_i) \to p_i$ where p_i are fixed values in the range $(0, 1)$, $i = 1, 2, \ldots, k$, then the distribution of X tends to the distribution of Y.*

Proof. Let $\lim_H$ denote limit as $\alpha \to \infty$, $\beta_i \to \infty$ so that $\beta_i/(\alpha + \Sigma_{i=1}^{k}\beta_i) \to p_i$, $i = 1, 2, \ldots, k$. Then

$$\lim_H P(X=x) = \sum_{\Sigma_{i=1}^{k} i x_i = x} \binom{n}{x_1, \ldots, x_k, n - \Sigma_{i=1}^{k} x_i}$$

$$\times \lim_H \left\{ \left(\frac{\alpha}{\alpha + \Sigma_{i=1}^{k}\beta_i} \right)^{n - \Sigma_{i=1}^{k} x_i} \prod_{i=1}^{k} \left(\frac{\beta_i}{\alpha + \Sigma_{i=1}^{k}\beta_i} \right)^{x_i} \right\}$$

$$= \sum_{\Sigma_{i=1}^{k} i x_i = x} \binom{n}{x_1, \ldots, x_k, n - \Sigma_{i=1}^{k} x_i} \left(1 - \sum_{i=1}^{k} p_i \right)^{n - \Sigma_{i=1}^{k} x_i} \prod_{i=1}^{k} p_i^{x_i}.$$

Hence the result. $\square$

3. The cluster negative hypergeometric distribution

Suppose now that we wish to sample n balls one at a time from the urn of the previous section, altering its composition at each stage by adding a ball of the same type as the sampled one before the next draw is made. As far as applications are concerned this allows for the hypothesis of contagion since each added ball increases the probability of sampling a ball of the same type at the next draw.

So we start off with an urn containing α balls marked 0 and β_i balls marked i, $i = 1, 2, \ldots, k$. A ball is drawn, its number is recorded and the ball is returned to the urn with one additional ball of the same type. Let X be the sum of the numbers observed in a sample of n balls sampled in the described manner. Then the following theorems can be shown following arguments similar to those used to prove Theorems 2.1 and 2.2.

Theorem 3.1. *The probability of the event $\{X = x\}$ is given by*

$$P(X=x) = \sum_{\Sigma_{i=1}^{k} i x_i = x} \binom{n}{x_1, \ldots, x_k, x - \Sigma_{i=1}^{k} x_i} \frac{(\beta_1)_{(x_1)} \cdots (\beta_k)_{(x_k)} \alpha_{(n - \Sigma_{i=1}^{k} x_i)}}{(\alpha + \Sigma_{i=1}^{k}\beta_i)_{(n)}}, \tag{3.1}$$

$x = 0, 1, 2, \ldots, nk$. $\square$

Theorem 3.2. *Let X be a random variable defined as in Theorem 3.1. Then $\Sigma_{x=0}^{nk} P(X = x) = 1$.* $\square$

The result of this theorem implies that (3.1) defines a probability distribution which for $k = 1$ coincides with the ordinary negative hypergeometric distribution.

Definition 3.1. Let n, k be fixed positive integers and let α, β_i, $i = 1, 2, \ldots, k$, be positive real numbers. A random variable X will be said to have the cluster negative hypergeometric distribution with parameters $k\,n$, α, $\beta_1, \ldots, \beta_k$ ($X \sim H_k^-(x;\ n,\ \alpha,\ \beta_1, \ldots, \beta_k)$) if its probability function is given by (3.1).

As before, if instead of the rational proportion $\beta_i/(\alpha + \Sigma_{i=1}^k \beta_i)$ of balls in the urn we start off with a proportion equal to any real number p_i in $(0, 1)$, the urn scheme considered in this section provides an alternative derivation of Steyn's (1956) univariate negative factorial multinomial distibution of occurrences of an event caused repeatedly by inversely sampled items of a population.

Theorem 3.3. *Let X be a non-negative integer-valued random variable with probability function as given by* (3.1). *Then*

(i) $$G_X(s) = \frac{\alpha_{(n)}}{\left(\alpha + \Sigma_{i=1}^k b_i\right)_{(n)}} F_D\left(-n;\ \beta_1,\ \beta_2, \ldots, \beta_k;\ 1 - \alpha - n;\ s,\ s^2, \ldots, s^k\right),$$ (3.2)

(ii) $$E(X) = \frac{n}{\alpha + \Sigma_{i=1}^k \beta_i} \sum_{i=1}^k i\beta_i,$$ (3.3)

(iii) $$V(X) = \frac{n\left(\alpha + \Sigma_{i=1}^k \beta_i + n\right)\left\{-2\Sigma_{i<j} ij\beta_i\beta_j + \Sigma_{i=1}^k i^2\beta_i\left(\alpha + \Sigma_{j \neq i}\beta_j\right)\right\}}{\left(\alpha + \Sigma_{i=1}^k \beta_i\right)^2\left(\alpha + \Sigma_{i=1}^k \beta_i + 1\right)}.$$ (3.4)

Proof. The proof is analogous to that of Theorem 2.3.   □

The two theorems that follow demonstrate the fact that the cluster negative hypergeometric distribution and the cluster binomial distribution relate to each other in the same manner as their ordinary forms (for $k = 1$) do. In particular, it is shown that (3.1) arises as a Dirichlet mixture of (2.6). Moreover (2.6) is shown to provide an appoximation to (3.1) for certain limiting values of its parameters.

Theorem 3.4. *Let X be a non-negative integer valued random variable having the cluster binomial distribution with probability function as given by* (2.6). *Suppose that the parameter vector $p = (p_1, p_2, \ldots, p_k)$ has itself a distribution with probability density function of the form*

$$f(p) = \frac{\Gamma\left(\Sigma_{i=1}^k \beta_i + \alpha\right)}{\Gamma(\alpha)\Pi_{i=1}^k\Gamma(\beta_i)} p_1^{\beta_1 - 1} \cdots p_k^{\beta_k - 1}\left(1 - \sum_{i=1}^k p_i\right)^{\alpha - 1},$$ (3.5)

α, p_i, $\beta_i > 0$, $i = 1, 2, \ldots, k$, $\Sigma_{i=1}^k p_i < 1$. *Then the resulting distribution of X is the cluster negative hyper-geometric distribution with probability function given by* (3.1).

Proof.

$$P(X = x) = \int_0^1 \cdots \int_0^1 P(X = x \mid p)f(p)\, dp_1 \cdots dp_k$$

$$= \frac{\Gamma\left(\Sigma_{i=1}^k \beta_i + \alpha\right)}{\Gamma(\alpha)\Pi_{i=1}^k\Gamma(\beta_i)} \int_0^1 \cdots \int_0^1 \sum_{\Sigma_{i=1}^k ix_i = x} \binom{n}{x_1, \ldots, x_k,\ n - \Sigma_{i=1}^k x_i}$$

$$\times \left(1 - \sum_{i=1}^k p_i\right)^{n + \alpha - \Sigma_{i=1}^k x_i - 1}\left\{\prod_{i=1}^k p_i^{x_i + \beta_i - 1}\right\} dp_1 \cdots dp_k$$

$$= \frac{\Gamma\left(\Sigma_{i=1}^{k}\beta_i + \alpha\right)}{\Gamma(\alpha)\Pi_{i=1}^{k}\Gamma(\beta_i)} \sum_{\Sigma_{i=1}^{k}ix_i = x} \binom{n}{x_1, \ldots, x_k, \, n - \Sigma_{i=1}^{k}x_i}$$

$$\times \frac{\left\{\Pi_{i=1}^{k}\Gamma(x_i + \beta_i)\right\}\Gamma\left(n + \alpha - \Sigma_{i=1}^{k}x_i\right)}{\Gamma\left(\Sigma_{i=1}^{k}\beta_i + n + \alpha\right)}.$$

Hence, the result. □

Theorem 3.5. *Let X be a random variable distributed according to the cluster negative hypergeometric distribution with probability function as given by* (3.1) *and let Y be another random variable having the cluster binomial distribution with probability function given by* (2.6). *Then, if $\alpha \to \infty$ and $\beta_i \to \infty$ so that $\beta_i/(\alpha + \Sigma_{i=1}^{k}\beta_i) \to p_i$, $0 < p_i < 1$, $i = 1, 2, \ldots, k$, the distribution of X tends to the distribution of Y.*

Proof. Let $\lim_H$ stand for limit as $\alpha \to \infty$, $\beta_i \to \infty$ so that $\beta_i/(\alpha + \Sigma_{i=1}^{k}\beta_i) \to p_i$, $i = 1, 2, \ldots, k$. Then

$$\lim_{H} P(X = x) = P(Y = x)$$

since

$$\lim_{H} \frac{\alpha_{(n - \Sigma_{i=1}^{k}x_i)}\Pi_{i=1}^{k}(\beta_i)_{(x_i)}}{\left(\alpha + \Sigma_{i=1}^{k}\beta_i\right)_{(n)}} = \lim_{H}\left(\frac{\alpha}{\alpha + \Sigma_{i=1}^{k}\beta_i}\right)^{n - \Sigma_{i=1}^{k}x_i} \prod_{i=1}^{k} \lim_{H}\left(\frac{\beta_i}{\alpha + \Sigma_{i=1}^{k}\beta_i}\right)^{x_i}$$

$$= \left(1 - \sum_{i=1}^{k}p_i\right)^{n - \Sigma_{i=1}^{k}x_i} \prod_{i=1}^{k}p_i^{x_i}.$$

Hence the theorem has been established. □

The sampling scheme considered in this section can be slightly altered so as to yield a more general type of distribution. In particular, if each sampled ball is returned to the urn along with c additional balls of the same type ($c \geqslant 1$) before the next ball is drawn then X, the sum of the n sampled numbers follows a distribution with a probability function given by

$$P(X = x) = \sum_{\Sigma_{i=1}^{k}ix_i = x} \binom{n}{x_1, \ldots, x_k, \, n - \Sigma_{i=1}^{k}x_i} \frac{(\alpha/c)_{(n - \Sigma_{i=1}^{k}x_i)}\Pi_{i=1}^{k}(\beta_i/c)_{(x_i)}}{\left((\Sigma_{i=1}^{k}\beta_i + \alpha)/c\right)_{(n)}}, \tag{3.6}$$

$\alpha, c, \beta_i > 0$, $i = 1, 2, \ldots, k$, $x = 0, 1, 2, \ldots, nk$. For $k = 1$ this reduces to the ordinary Polya distribution.

Definition 3.2. A non-negative, integer-valued random variable X taking values in $\{0, 1, \ldots, nk\}$, where k is a positive integer is said to have the cluster Polya distribution with parameters c, α and β_i, $i = 1, 2, \ldots, k$, if its probability function is given by (3.6).

Note that when $c = 0$, (3.6) reduces to the cluster binomial as this is defined by (2.6) for $p_i = \beta_i/(\alpha + \Sigma_{i=1}^{k}\beta_j)$, $i = 1, 2, \ldots, k$. Moreover when $c = -1$, (3.6) reduces to the cluster hypergeometric with probability function as given by (2.1).

Hence (3.6) unifies the derivation of the distributions that arise in the context of direct cluster sampling, namely the cluster binomial of Panaretos and Xekalaki (1986a) and the cluster hypergeometric and the cluster negative hypergeometric distributions introduced in the present paper.

Acknowledgement

References

Panaretos, J. and E. Xekalaki (1986a), The stuttering generalized Waring distribution, *Statist. Probab. Lett.* **4**, 313–318.

Panaretos, J. and E. Xekalaki (1986b), On generalized binomial and multinomial distributions and their relation to generalized Poisson distributions, *Ann. Inst. Statist. Math. A* **38**, 223–231.

Polya, G. (1931), Sur quelques points de la theorie des probalitées, *Ann. Inst. H. Poincaré* **1**, 117–161.

Steyn, H.S. (1956), On the univariable series $F(t) = F(a; b_1, b_2, \ldots, b_k; c; t, t^2, \ldots, t^k)$ and its application in probability theory, *Proc. Kon. Ned. Akad. Wetensch. Ser. A* **59**, 190–197.

Xekalaki, E. and J. Panaretos (1989), On some distributions arising in inverse cluster sampling, *Comm. Statist. A—Theory Methods* **18** (1) 355–366.